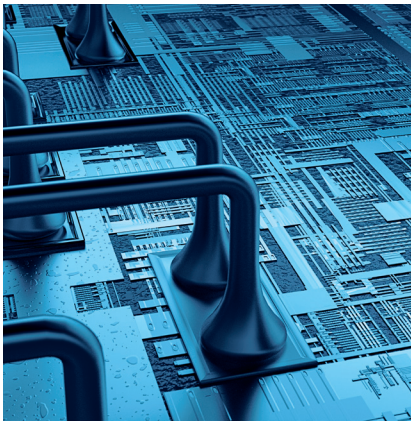
CASE STUDY
Intel® Ethernet 10 Gigabit Server Adapters
Single-Root I/O Virtualization
NASA

NASA's Flexible Cloud Fabric:
Moving Cluster Applications to the Cloud



The NASA Center for Climate Simulation found that an open cloud architecture using 10 Gigabit Ethernet for both inter-node communication and management traffic is a viable alternative to its purpose-built InfiniBand*-based cluster for many large-scale modeling applications. The organization hopes to capture the elasticity and flexibility benefits of both cloud computing and converged networking on Ethernet.

Network architects at the NASA Center for Climate Simulation recently began investigating the viability of running the organization's modeling and simulation applications on cloud infrastructure, as an alternative to its purpose-built computing cluster named Discover. Hoping to capture the inherent advantages of cloud infrastructure, such as agility and elasticity, they want to establish whether an open cloud architecture can meet the applications' rigorous throughput and latency requirements. In particular, they need to ensure that overhead associated with virtualization would not limit performance.

As part of the shift to the cloud, the team hopes to converge the environment's backbone and management infrastructures onto 10 Gigabit Ethernet. Using a single network fabric is expected to help optimize the flexibility and cost-effectiveness of the overall solution.

Traditional Architecture for Cluster Computing

The NASA Center for Climate Simulation's research on climate change and related phenomena, which requires extensive computer modeling, contributes to efforts such as hurricane prediction, analysis of past weather patterns, and scientific support of government climate policy. The cluster named Discover that has done this work for some years uses an integrated set of supercomputing, visualization, and data-management technologies to deliver roughly 400 teraflops of capacity:

- **Compute resources**: 30,000 conventional Intel® Xeon® processor cores and 64 GPUs
- **Inter-node backbone**: DDR and QDR InfiniBand*
- **Management networking**: Gigabit and 10 Gigabit Ethernet (GbE and 10GbE)
- **Data store**: ~4 petabyte RAID-based parallel file system (GPFS), plus ~20 petabyte tape archive

Discover is based entirely on non-virtualized machines, so adding capacity requires additional physical servers to be provisioned. Reducing the traditional cost and complexity of those changes is one benefit of cloud computing. Moreover, cloud architectures add elasticity that aids in job scheduling and helps avoid operational bottlenecks associated with long-running jobs.

The Nebula Cloud: A Proposed Adjunct to the Existing Cluster Approach

NASA's Nebula is a cloud-based infrastructure-as-a-service environment, conceived as an alternative means of delivering compute capacity without the costly construction of additional data centers. There are currently two Nebula environments, which have been built in modular freight containers; one is deployed at the Goddard Space Flight Center in Maryland, and the other is at Ames Research Center in California.

Nebula is based on OpenStack*, an open source software project that enables organizations to build private and public clouds. Backed by more than 100 companies and 1,000 developers, OpenStack is designed to support massively scalable cloud computing infrastructures. Intel is actively involved[1] in the project, helping enable OpenStack to take advantage of Intel® platform features.

The key question surrounding the viability of Nebula as an alternative to Discover is whether it can deliver equivalent performance. In particular, the team must determine whether the virtualized environment on which Nebula is based will introduce overhead or other factors that will create unacceptable limitations, compared to "bare-metal" clusters.

Test Methodology for Moving Applications to the Cloud

To meet critical speed and latency requirements in node-to-node communication, NASA performance engineers worked with Intel to employ virtualization technologies to their full potential. Together, the team established a test methodology to compare the two environments on several workloads, including the Nuttcp network performance measurement tool[2], the Ohio State University MPI Benchmarks[3], and the Intel® Math Kernel Library (Intel® MKL) implementation of LINPACK[4]. Analysis using these benchmarks enabled the team to measure and compare system throughput and latency between various types of physical or virtual servers:

- **Bare-metal**. Data transfer between non-virtualized servers.
- **Software-only virtualization**. Data transfer between virtual machines (VMs).
- **Virtualized I/O**. Data transfer between VMs with OS-based paravirtualization.
- **Single-Root I/O Virtualization (SR-IOV)**. Data transfer between VMs using SR-IOV.

The test systems were Dell PowerEdge* R710 servers, configured as shown in Table 1.

As mentioned above, the purpose of this test methodology was to determine whether the Nebula cloud infrastructure using 10GbE can deliver throughput and latency equivalent to that of the Discover cluster using InfiniBand. More specifically, the approach of comparing multiple virtualization scenarios enabled the testing to reveal the role those virtualization technologies can play in meeting performance goals.

Results: SR-IOV is a Key Requirement for Moving Cluster Applications to the Cloud

The set of test results based on the Nuttcp benchmark is shown in Table 2; similar to the better-known Netperf benchmark, Nuttcp measures raw network bandwidth. In the table, the figures for individual trials are arranged from the top row to the bottom starting with the lowest throughput measure for each test scenario, through the highest. In the testing of data transfer from one bare-metal server to another (the leftmost column), the highest throughput rates come fairly near the wire speed of the 10GbE port.

The second column of Table 2 shows dramatic decay in throughput for the software-only virtualization case, with rates falling to just a few percent of full

Table 1. Test-bed configuration.

	Bare-metal	VMs
Processors	Two Intel® Xeon® processors E5520 @ 2.27 GHz (quad-core)	
Main memory	48 GB	16 GB
OS	Ubuntu* Linux* 11.04 (Kernel 2.6.38-10.server)	
Hypervisor	NA	NA

Table 2. Nuttcp results, which demonstrate that SR-IOV helps attain virtualized throughput near wire speed, similar to that with bare-metal servers.

Bare Metal-to-Bare Metal	VM-to-VM (Software Virtualization)	VM-to-VM (Virtualized I/O)	VM-to-VM (with SR-IOV)
4418.8401 Mbps	137.3301 Mbps	5678.0625 Mbps	8714.4063 Mbps
8028.6459 Mbps	138.5963 Mbps	5692.8146 Mbps	8958.5032 Mbps
9341.4362 Mbps	141.8702 Mbps	5746.2926 Mbps	9101.7356 Mbps
9354.0999 Mbps	145.6024 Mbps	5864.0557 Mbps	9151.5769 Mbps
9392.7072 Mbps	145.7500 Mbps	5955.8176 Mbps	9193.1103 Mbps
9414.7318 Mbps	146.1043 Mbps	5973.2256 Mbps	9228.5370 Mbps
9414.8207 Mbps	146.1092 Mbps	6223.4034 Mbps	9251.8453 Mbps
9414.9368 Mbps	146.2758 Mbps	6309.8478 Mbps	9313.8894 Mbps
9415.1618 Mbps	146.3042 Mbps	6311.3896 Mbps	9348.2984 Mbps
9415.2675 Mbps	146.4449 Mbps	6316.7924 Mbps	9408.0323 Mbps

wire speed. Clearly, this software-only virtualization configuration is insufficient to support the high-performance computing demands of the NASA Center for Climate Simulation. On the other hand, the figures in the third column show that virtualizing I/O with the help of hardware acceleration drives throughput up considerably, although the highest throughput figures achieved in this test case are less than 65 percent of wire speed.

The rightmost column of Table 2 shows dramatic throughput improvement in the virtualized environment when SR-IOV is utilized. In fact, the figures in this column approach those of the bare-metal case, indicating that a properly configured virtualized network can deliver throughput that is roughly equivalent to that of a non-virtualized one.

To expand on the Nuttcp results, the test team performed trials on the other two benchmarks with different message sizes. Figure 1 shows throughput (left chart) and latency (right chart) results for the Ohio State MPI benchmark. Surprisingly, the test configuration that uses SR-IOV actually outperforms the bare-metal one. The test team postulates that this performance differential is due to inefficiencies in the Linux* kernel that are overcome by direct assignment under SR-IOV. In any event, this test result does support the finding above that, in some cases, virtualized performance with SR-IOV can be comparable to equivalent non-virtualized performance.

Finally, the test team considered the results of throughput testing with the Intel MKL implementation of LINPACK, as shown in Figure 2. Here, while the SR-IOV implementation increases performance relative to the non-SR-IOV case, its performance is somewhat lower than

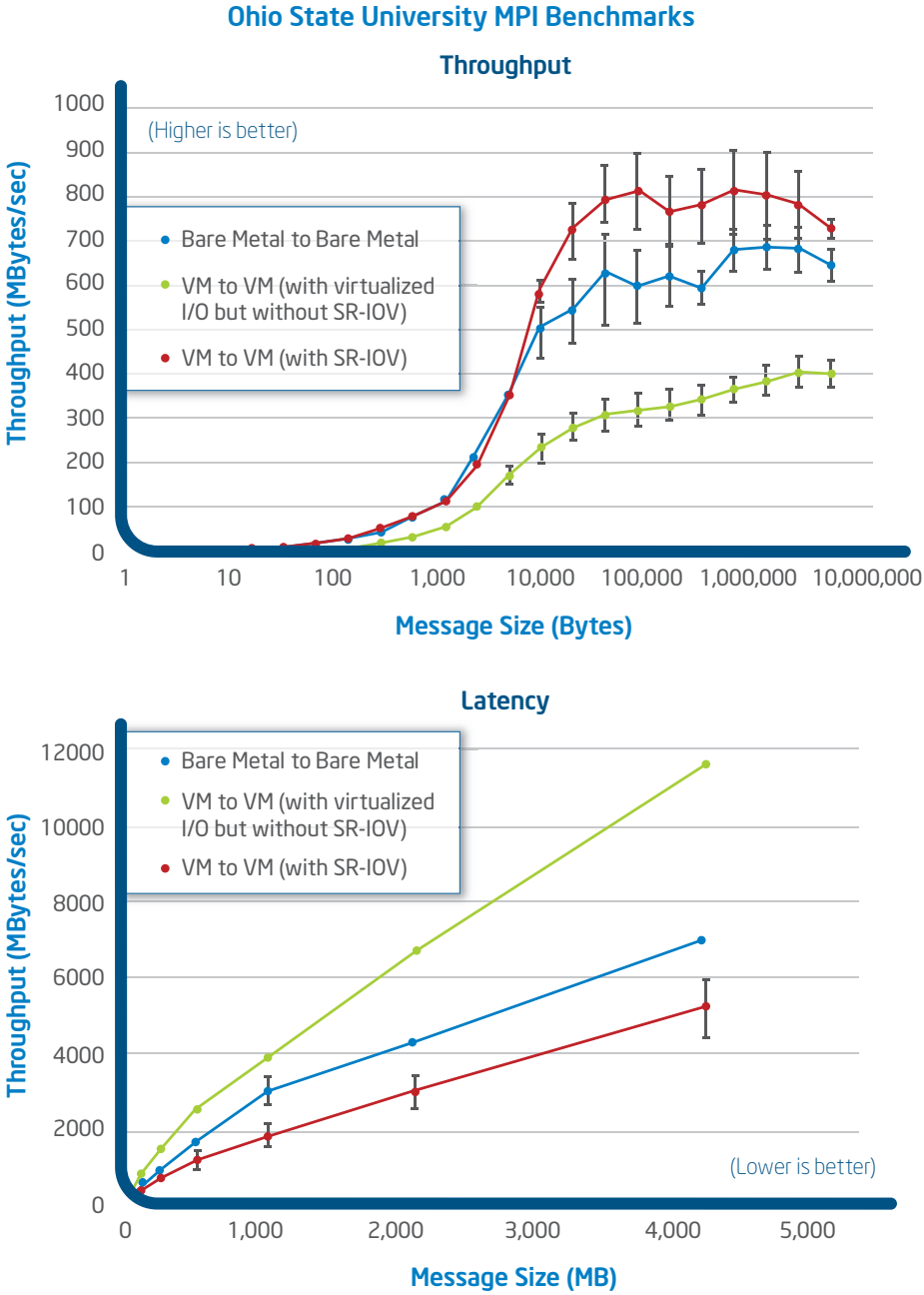


Figure 1. Virtualized and non-virtualized performance results for the Ohio State University MPI benchmark.

SINGLE-ROOT I/O VIRTUALIZATION (SR-IOV) DEFINED

Supported by Intel® Ethernet Server Adapters, SR-IOV is a standard mechanism for devices to advertise their ability to be simultaneously shared among multiple virtual machines (VMs). SR-IOV allows for the partitioning of a PCI function into many virtual functions (VFs) for the purpose of sharing resources in virtual or non-virtual environments. Each VF can support a unique and separate data path for I/O-related functions, so for example, the bandwidth of a single physical port can be partitioned into smaller slices that may be allocated to specific VMs or guests.

(although in the same realm as) the bare-metal case. This result indicates that there are some cases where bare-metal (non-virtualized) implementations will continue to outperform virtualized ones. Therefore, while virtualized environments are viable for some high-performance workloads, the performance and latency requirements of individual applications must be considered.

Conclusions

The core conclusion from this testing is that cloud-based high-performance computing is a viable possibility. SR-IOV, as supported by Intel Ethernet Server Adapters, is a core enabling technology that helps overcome performance limitations associated with virtualization.

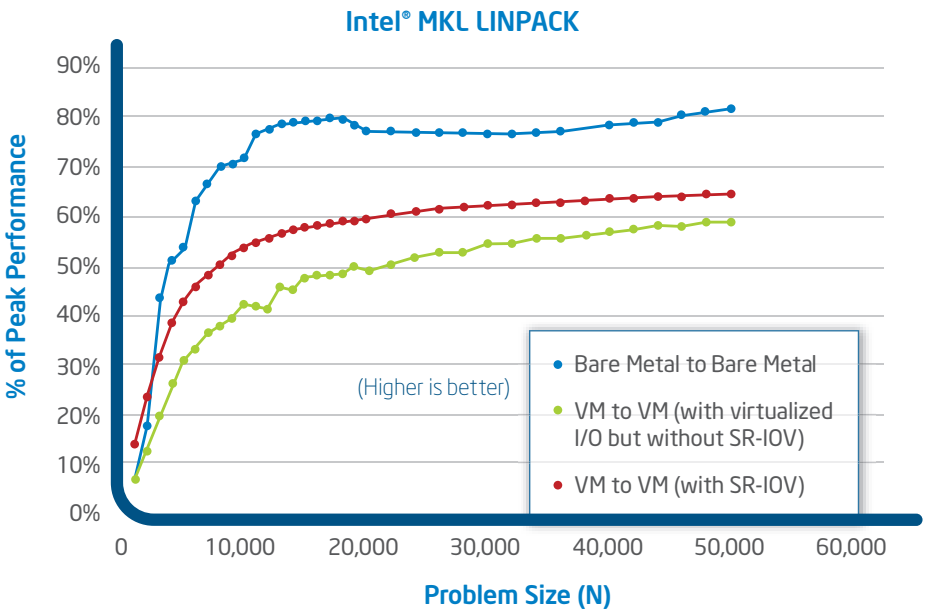


Figure 2. Virtualized and non-virtualized performance results for Intel® MKL LINPACK.

In terms of both throughput and latency test results using the Ohio State University MPI Benchmark, virtualized performance actually exceeded that of the bare metal (non-virtualized) case. While in testing with the Intel MKL LINPACK implementation, bare-metal servers out-performed virtualized ones, the benefit delivered by SR-IOV nevertheless demonstrates the potential for virtualized usage models. Because of the flexibility and scalability afforded by virtualization (including the cloud), these results merit further consideration.

Future Work

To advance the state of this preliminary testing, additional work is needed. In particular, the team wants to test additional benchmarks and real-world applications, as well as extending the tests to include InfiniBand fabric and cloud infrastructures such as OpenStack and Eucalyptus*. Future testing will also include additional hypervisors, such as Xen* and other VM OSs, such as Red Hat Enterprise Linux and SUSE Linux.

As NASA continues to refine its cloud-based infrastructure as a service, it expects to realize more benefits in the areas of simplification, flexibility, and cost-effectiveness. Looking ahead, the agency's high-performance computing workloads have begun the process of shifting to open infrastructures that use Ethernet fabric, and further acceleration seems inevitable.

SOLUTION PROVIDED BY:

For more information about the Nebula project and service, visit **http://nebula.nasa.gov**

For more information about the NASA Center for Climate Simulation, visit **http://www.nccs.nasa.gov**

For more information about Intel® Ethernet, visit **www.intel.com/go/ethernet**

[1] http://software.intel.com/sites/oss/project/openstack.php.

[2] Nuttcp-7.1.5.c (gcc compiler): http://lcp.nrl.navy.mil/nuttcp.

[3] MVAPICH2 1.7rc1 (Intel® compiler): http://mvapich.cse.ohio-state.edu/.

[4] Intel® MKL 10.2.6 (Intel compiler): http://software.intel.com/en-us/articles/intel-math-kernel-library-linpack-download/.

Printed in USA 1111/BY/MESH/PDF ♻ Please Recycle 326332-001US